

A comparative analysis of the performance of large language models in the basic life support exam: Comprehensive evaluation of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1

LLMs' performance in basic life support exam

Bensu Bulut¹, Medine Akkan Öz¹, Murat Genç², Ayşenur Gür³, Mehmet Yortanlı⁴, Betül Çiğdem Yortanlı⁵, Ramiz Yazıcı⁶, Hüseyin Mutlu⁷, Mustafa Sirri Kotanoglu⁸, Eray Cinar⁹, Ramazan Kocaaslan¹⁰

¹ Department of Emergency Medicine, Health Science University, Ankara Gulhane Training and Research Hospital, Ankara

² Department of Emergency Medicine, Ankara Training and Research Hospital, Ankara

³ Department of Emergency Medicine, Etimesgut Şehit Sait Ertürk State Hospital, Ankara

⁴ Department of Emergency Medicine, Konya Numune Hospital, Konya

⁵ Department of Internal Medicine, University of Health Sciences, Konya City Hospital, Konya

⁶ Department of Emergency Medicine, Health Science University, Istanbul Kanuni Sultan Suleyman Training and Research Hospital, Istanbul

⁷ Department of Emergency Medicine, Aksaray University, Aksaray Training and Research Hospital, Aksaray

⁸ Department of Anesthesiology And Reanimation, Ankara Training and Research Hospital, Ankara

⁹ Department of Thoracic Surgery, University of Health Sciences, Bilkent City Hospital, Ankara

¹⁰ Department of Urology, Kafkas University, Kars, Turkey

Abstract

Aim: Considering the growing role artificial intelligence technologies play in medical education, this study aims to provide a comparative evaluation of the performances of large language models ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 in the Basic Life Support (BLS) Exam.

Materials and Methods: In this observational study, we presented four large language models with 25 multiple-choice questions based on the American Heart Association (AHA) guidelines. Questions were divided into two categories as knowledge-based (n = 14, 56%) and case-based (n = 11, 44%). Response consistency was ensured by presenting each question on three separate days to all models. Models' accuracy rates were assessed using overall accuracy, strict accuracy, and ideal accuracy criteria.

Results: In the overall accuracy assessment, ChatGPT-4o and DeepSeek R1 models showed 100% success, and Gemini 2.0 and Claude 3.5 models achieved 96% success rate. All models performed perfectly on the case-based questions. On the knowledge-based questions, ChatGPT-4o and DeepSeek R1 scored full points, while Gemini 2.0 and Claude 3.5 achieved 90.9% success. Statistical analysis showed no significant difference between results (p = 0.368).

Discussion: Large language models show high accuracy rates in BLS training. These technologies can be used in supportive roles in medical education, but human supervision is critical in clinical decision-making.

Keywords

Artificial Intelligence, Large Language Models, Basic Life Support, Medical Education, ChatGPT, Resuscitation

DOI: 10.4328/ACAM.22758 Received: 2025-05-28 Accepted: 2025-07-29 Published Online: 2025-07-30 Printed: 2025-08-01 Ann Clin Anal Med 2025;16(8):578-581

Corresponding Author: Bensu Bulut, Department of Emergency Medicine, Health Science University, Ankara Gulhane Training and Research Hospital, Ankara, Turkey.

E-mail: bensu.bulut@gmail.com P: +90 553 181 93 62

Corresponding Author ORCID ID: <https://orcid.org/0000-0002-6320-9667>

Other Authors ORCID ID: Medine Akkan, <https://orcid.org/0000-0002-6320-9667> · Murat Genç, <https://orcid.org/0000-0003-3407-1942>

Ayşenur Gür, <https://orcid.org/0000-0002-9521-1120> · Mehmet Yortanlı, <https://orcid.org/0000-0002-6744-2423> · Betül Çiğdem Yortanlı, <https://orcid.org/0000-0003-2698-3159>

Ramiz Yazıcı, <https://orcid.org/0000-0001-9210-914X> · Hüseyin Mutlu, <https://orcid.org/0000-0002-1930-3293> · Mustafa Sirri Kotanoglu, <https://orcid.org/0000-0002-6906-573X>

Eray Cinar, <https://orcid.org/0000-0002-4564-6097> · Ramazan Kocaaslan, <https://orcid.org/0000-0003-1944-7059>

Introduction

Basic Life Support (BLS) is a critical procedure that includes timely initiation of life-saving interventions in life-threatening situations such as cardiac or respiratory arrest. Timely and appropriate resuscitation significantly improves survival rates and neurological outcomes of patients [1]. Organisations such as the American Heart Association (AHA) organise and deliver courses around the world to improve accessibility of cardiopulmonary resuscitation (CPR) and automatic external defibrillators (AED) training [2]. Similarly, the Resuscitation Council UK (RCUK) organizes Basic Life Support (BLS) and Advanced Life Support (ALS) courses that are mandatory for all healthcare providers, and these courses are updated every five years [3]. In all these courses, pre-course tests are performed to facilitate participant learning, and post-course tests are performed to provide certification for participants.

The rapid development of artificial intelligence (AI) technologies is leading to important transformations in medical education. Nowadays, large language models (LLMs) are playing ever-increasing roles in the training of healthcare professionals through platforms such as ChatGPT, Bard, and Claude [4]. Studies focusing on the performance of large language models in medical exams have increased remarkably in recent years [4,5,6]. King et al.'s study, where GPT-4V achieved success rates of 96% in BLS and 90% in ACLS exams, shows the potential of these technologies in medical education [7]. In their study where they performed a comparative analysis of the performances of ChatGPT and Bard in resuscitation-based medical questions, Patel et al. found similar results where these models achieved success rates between 75-96% [4]. These successful performances suggest that AI models can be used effectively as supportive tools in medical education [8]. It has been reported that AI-supported tools can provide high-quality and empathetic health recommendations, and in some cases, can even surpass physicians' answers [9].

However, despite the successful performances of specifically engineered software, the question of whether widely used general-purpose chatbots such as GPT-3.5, GPT-4, Bard, and Bing have the same competence in emergencies is still unanswered. The study by Aqavil-Jahromi et al. revealed that there are significant differences in the accuracy and reliability of different AI models in BLS scenarios [10]. Meanwhile, Birkun's study showed that AI-based chatbots do not generate content in accordance with resuscitation guidelines and can give potentially harmful recommendations [11].

The aim of this study is to analyse the performances of four large language models, such as ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 in standard questions based on the American Heart Association's (AHA) BLS protocols and the reliability of these chatbots in emergencies. In our study, all models were asked the same BLS questions on three separate days, and the answers were assessed using the overall accuracy, strict accuracy, and ideal accuracy criteria. The findings of this study have critical importance for understanding the potentials and limitations of large language models in emergencies, parallel to technological advancements.

Materials and Methods

In this observational study, response accuracy and consistency

of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 models' performances in answering Basic Life Support (BLS) module questions were examined. Models' comparative performances were analysed in overall accuracy, strict accuracy, and ideal accuracy levels. BLS training has been a routine training provided to all healthcare workers and even students for many years, and is updated every five years [1,2]. The BLS exam consists of 25 multiple-choice questions with five options. BLS courses are the highest level courses based on the American Heart Association (AHA) BLS guidelines [3].

We used four large language models (LLMs) in our study. First is ChatGPT-4o; ChatGPT 4 Omni (ChatGPT-4o) version of ChatGPT, which is currently known to have the highest level of medical knowledge among its peers, was used [4]. This model is trained on extensive data, including medical texts and journals published up to and including September 2024 [7,9]. Second is Gemini 2.0, representing Google's evolving family of large language models (including Bard and earlier Gemini versions), known for producing medically relevant, accurate, and referenced outputs in scientific contexts [10,12]; third is Claude 3.5, used for its accuracy and quality levels similar to that of physicians in the medical field [13]; fourth is DeepSeek R1, used because it is reported as a promising AI tool in facilitating the diagnosis of diseases and conditions [14,15].

25 multiple choice BLS questions were assessed by authors R.Y. and H.M. separately and divided into two groups as knowledge-based questions and case-based questions. Where the assessment of these authors differed, questions were assessed by author B.B. for a final decision. Out of all BLS questions, 11 (44%) were case-based and 14 (56%) were knowledge-based. BLS questions were presented to ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 models once each on three separate days between 1-30 January, and three answers were generated for each question. This approach is similar to other studies where LLMs are presented with a question three times in order to provide LLM consistency and response stability [4,14]. Models' accuracy rates were evaluated using overall accuracy, strict accuracy, and ideal accuracy criteria.

Overall accuracy: Considered correct when all three responses are correct.

Strict accuracy: Considered correct when two out of three responses are correct.

Ideal accuracy: Considered correct when one out of three responses is correct.

AHA's BLS questions and answers and responses of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 models were recorded in a separate Microsoft Excel 2023 file (Version 16.73, Microsoft Corporation, Redmond, WA). Because only artificial intelligence models were used in this study, and there were no human or animal test subjects, it did not require ethics committee approval.

Statistical Method

Statistical analysis was performed using SPSS 27.0 (IBM Corp, Armonk, NY, USA). Categorical variables were expressed as numbers and percentages (%). Categorical variables were analysed by the Chi-square and Fisher's exact test. Models' performances were compared using the Cochran Q test. Statistical differences between the models were evaluated using the McNemar test for pairwise comparisons. Fleiss'

Kappa test was applied to measure the consistency between the responses of models in three separate sessions.

Results

BLS exam questions were presented to four large language models (LLMs): ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1. All models showed complete success in case-based questions. On knowledge-based questions, ChatGPT-4o and DeepSeek R1 achieved 100% accuracy, while Gemini 2.0 and Claude 3.5 models achieved 90.9% accuracy rates. According to strict accuracy criteria, ChatGPT-4o, Gemini 2.0, and DeepSeek

R1 achieved %100 success, while Claude 3.5 was inconsistent on one question and achieved only %96 accuracy. On the ideal accuracy level, only Claude 3.5 lagged behind other models with an accuracy rate of 96%. There was no significant difference between models in the statistical analysis ($p = 0.368$) (Table 1, Figure 1, Figure 2).

Discussion

In this study, we compared the performances of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 models in Basic Life Support (BLS) questions. Our study is the first study that evaluates the performances of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 in the BLS exam. Our findings demonstrate the increasing role AI models play in medical education, while also highlighting the current limitations. While ChatGPT-4o and DeepSeek R1 perform with %100 accuracy on the overall accuracy measure reflects these models’ perfect progress in medical knowledge, the 96% accuracy rate of Gemini 2.0 and Claude 3.5 is also satisfactory. Similarly, in their study comparing the performances of ChatGPT and Bard in BLS exams, Patel et al. reported that Bard achieved 96% accuracy rates, and ChatGPT’s performance was 75% [4]. These findings show that large language models can be utilised effectively as medical training tools.

In the study of Fijačko et al., ChatGPT’s performance in AHA exams differed between 68% and 92.1% according to the model used [16]. Patel et al. reported that ChatGPT showed success rates ranging between 75-96% for medical questions [4]. Kokulu et al. reported that ChatGPT achieved 94% success in the Pediatric Advanced Life Support (PALS) exam [17]. Similarly, in their study evaluating the performance of GPT-4 in BLS and ACLS exams, King et al. showed a success rate of 96% in the BLS exam [7]. In our study, we found that ChatGPT-4o and DeepSeek R1 achieved % 100% accuracy rates in knowledge-based questions. The high success rates achieved in our study demonstrate the perfect progress of integrating AI models into medical education and highlight the potential of these technologies.

The complete success shown by our models in case-based questions reflects their potential in analysing clinical scenarios. The fact that all models achieved 100% accuracy rates in case-based questions demonstrates that these tools can be used in practical resuscitation training. Beck et al. also obtained similar findings and concluded that AI models could be useful in clinical case analysis [18]. However, this success needs to be interpreted with caution as there are more complex variables in real clinical practice.

One of the remarkable findings of our study was that Claude 3.5 lagged behind other models in case-based questions. Claude 3.5 performing with an accuracy rate of 90.9% in case-based questions demonstrates that different models perform differently in complex clinical scenarios. Beck et al. also obtained similar findings in their study, where they evaluated the European Council (ERC) guidelines with different ChatGPT versions. In this study, ChatGPT-3.5 was reported to be only 77% compliant with the guidelines, whereas ChatGPT-4 was 84% compliant [18]. These findings indicate that while AI models are competent in the transfer of medical knowledge,

Table 1. Accuracy comparison of large language models

	ChatGPT 4o	Gemini 2.0	Claude 3.5	DeepSeek R1
	n (%)	n (%)	n (%)	n (%)
Overall accuracy				
Total (n = 25)	25 (100.0)	24 (96.0)	24 (96.0)	25 (100.0)
Case (n = 11)	11 (100.0)	10 (90.9)	10 (90.9)	11 (100.0)
Knowledge (n = 14)	14 (100.0)	14 (100.0)	14 (100.0)	14 (100.0)
Strict accuracy				
Total (n = 25)	25 (100.0)	25 (100.0)	24 (96.0)	25 (100.0)
Case (n = 11)	11 (100.0)	11 (100.0)	10 (90.9)	11 (100.0)
Knowledge (n = 14)	14 (100.0)	14 (100.0)	14 (100.0)	14 (100.0)
Ideal accuracy				
Total (n = 25)	25 (100.0)	25 (100.0)	24 (96.0)	25 (100.0)
Case (n = 11)	11 (100.0)	11 (100.0)	10 (90.9)	11 (100.0)
Knowledge (n = 14)	14 (100.0)	14 (100.0)	14 (100.0)	14 (100.0)

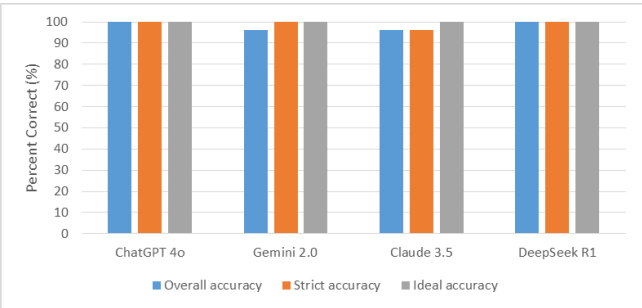


Figure 1. Performance comparison of large language models

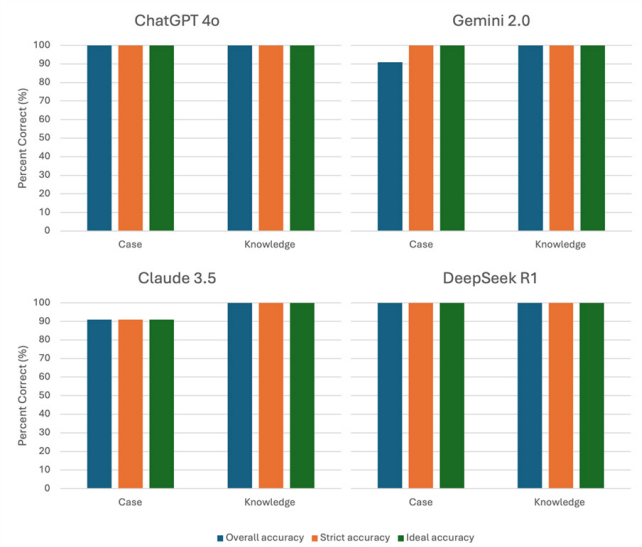


Figure 2. Model performance by question type

there are still areas open for improvement when analysing complex clinical scenarios.

Providing a comprehensive assessment of four different large language models and being one of the first comprehensive studies to include different question types (case and knowledge-based) are the strengths of our study. Also, presenting each question three times in order to assess response consistency improves the reliability of our findings. However, our study also has weaknesses. Firstly, we evaluated only 25 BLS questions, and 14 of these were knowledge-based and 11 were case-based. A larger question pool could better evaluate the models' performances with questions of different difficulty levels. Secondly, our study was limited to four large language models, and other available AI models were not included. Finally, due to the constant updating of language models, our findings are valid only for a specific version, and performance may change with future updates.

Conclusion

This study comprehensively evaluated the performances of large language models in Basic Life Support training and provided important findings. The performances of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1 in BLS questions were overall satisfactory. In particular, ChatGPT-4o and DeepSeek R1 achieve 100% accuracy rates in knowledge-based questions, indicates that these models can be used as reliable sources of information in medical education. However, their current technological levels are not mature enough to provide services as independent advisors in clinical settings. At this time, when human supervision and oversight are necessary, the use of these technologies as training and supporting tools should be considered, but they should not be relied upon to make vital decisions. Future studies should evaluate the adaptive capacity of these models to updated guidelines and their performance in multi-language education.

Scientific Responsibility Statement

The authors declare that they are responsible for the article's scientific content including study design, data collection, analysis and interpretation, writing, some of the main line, or all of the preparation and scientific review of the contents and approval of the final version of the article.

Animal and Human Rights Statement

All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Funding: None

Conflict of Interest

The authors declare that there is no conflict of interest.

References

1. Blewer AL, Ibrahim SA, Leary M, et al. Cardiopulmonary resuscitation training disparities in the United States. *J Am Heart Assoc.* 2017;6(5):e006124.
2. Bray JE, Smith K, Case R, Cartledge S, Stranely L, Finn J. Public cardiopulmonary resuscitation training rates and awareness of hands-only cardiopulmonary resuscitation: a cross-sectional survey of Victorians. *Emerg Med Australas.* 2017;29(2):158-64.
3. Soar J, Böttiger BW, Carli P, et al. European Resuscitation Council Guidelines 2021: Adult advanced life support. *Resuscitation.* 2021;161:115-51.
4. Patel S, Patel R. Embracing Large language models for adult life support learning. *Cureus.* 2024;16(12):e75961.
5. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med.* 2023;3(1):141.
6. Mutlu H, Kokulu K, Sert ET, Topuz MA. Evaluation of ChatGPT's performance in Türkiye's first emergency medicine sub-specialization exam. *Eur J Emerg Med.* 2025;24(1):17-26.

7. King RC, Bharani V, Shah K, Yeo YH, Samaan JS. GPT-4V passes the BLS and ACLS examinations: An analysis of GPT-4V's image recognition capabilities. *Resuscitation.* 2024;195:110106.
8. Onan A, Simsek N, Elcin M, Turan S, Erbil B, Deniz KZ. A review of simulation-enhanced, team based cardiopulmonary resuscitation training for undergraduate students. *Nurse Educ Pract.* 2017;27:134-43.
9. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183(6):589-96.
10. Aqavil Jahromi S, Eftekhari M, Akbari H, Aligholi Zahraie M. Evaluation of correctness and reliability of GPT, Bard, and Bing chatbots' responses in basic life support scenarios. *Sci Rep.* 2025;15(1):11429.
11. Birkun AA, Gautam A. Large language model-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehosp Disaster Med.* 2023;38(6):757-63.
12. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: A comparative study of OpenAI's GPT 4 and Google's Gemini. *Comput Biol Med.* 2025;185:109545.
13. Jin H, Guo J, Lin Q, Wu S, Hu W, Li X. Comparative study of Claude 3.5 Sonnet and human physicians in generating discharge summaries for patients with renal insufficiency: assessment of efficiency, accuracy, and quality. *Front Digit Health.* 2024;6:1456911.
14. Curtin LB, Finn LA, Czosnowski QA, Whitman CB, Cawley MJ. Computer-based simulation training to improve learning outcomes in mannequin-based simulation exercises. *Am J Pharm Educ.* 2011;75(6):113.
15. Faray de Paiva L, Luijten G, Puladi B, Egger J. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: A historical control study. *Int J Surg.* 2025;110(6):123456.
16. Fijačko N, Gosak L, Štiglic G, Picard CT, Douma MJ. Can ChatGPT pass the life support exams without entering the American Heart Association course? *Resuscitation.* 2023;185:109732.
17. Kokulu K, Demirtaş MS, Sert ET, Mutlu H. ChatGPT and pediatric advanced life support: A performance evaluation. *Resuscitation.* 2024;205:110451.
18. Beck S, Kühner M, Haar M, Daubmann A, Semmann M, Kluge S. Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: A comparative analysis between the ERC 2021 guidelines and both ChatGPT 3.5 and 4. *Scand J Trauma Resusc Emerg Med.* 2024;32(1):95.

How to cite this article:

Bensu Bulut, Medine Akkan Öz, Murat Genç, Ayşenur Gür, Mehmet Yortanlı, Betül Çiğdem Yortanlı, Ramiz Yazıcı, Hüseyin Mutlu, Mustafa Sirri Kotanoglu, Eray Çinar, Ramazan Kocaaslan. A comparative analysis of the performance of large language models in the basic life support exam: Comprehensive evaluation of ChatGPT-4o, Gemini 2.0, Claude 3.5, and DeepSeek R1. *Ann Clin Anal Med* 2025;16(8):578-581